Chiem Exercise 5

MAT 115

Exercise #5

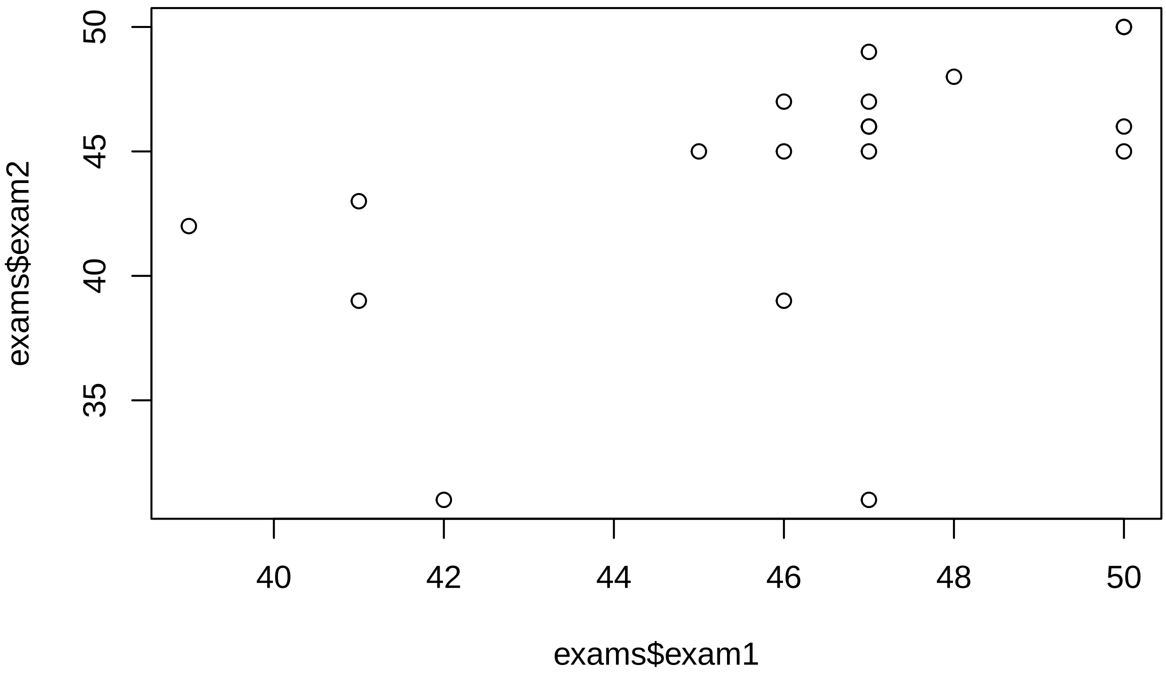
In this exercise we will introduce some basic plots available in base R (Section 2.15 of your text). We will get into fancier plots later, using various packages specifically meant to generate good-looking plots. For now, we look at just three types of statistical plots: scatterplots, histograms, and boxplots.

First, we load the dslabs package and the exams dataset.

```
library(dslabs)
load("exams.rda")
```

You can make a scatterplot of two quantitative variables by using the plot command:

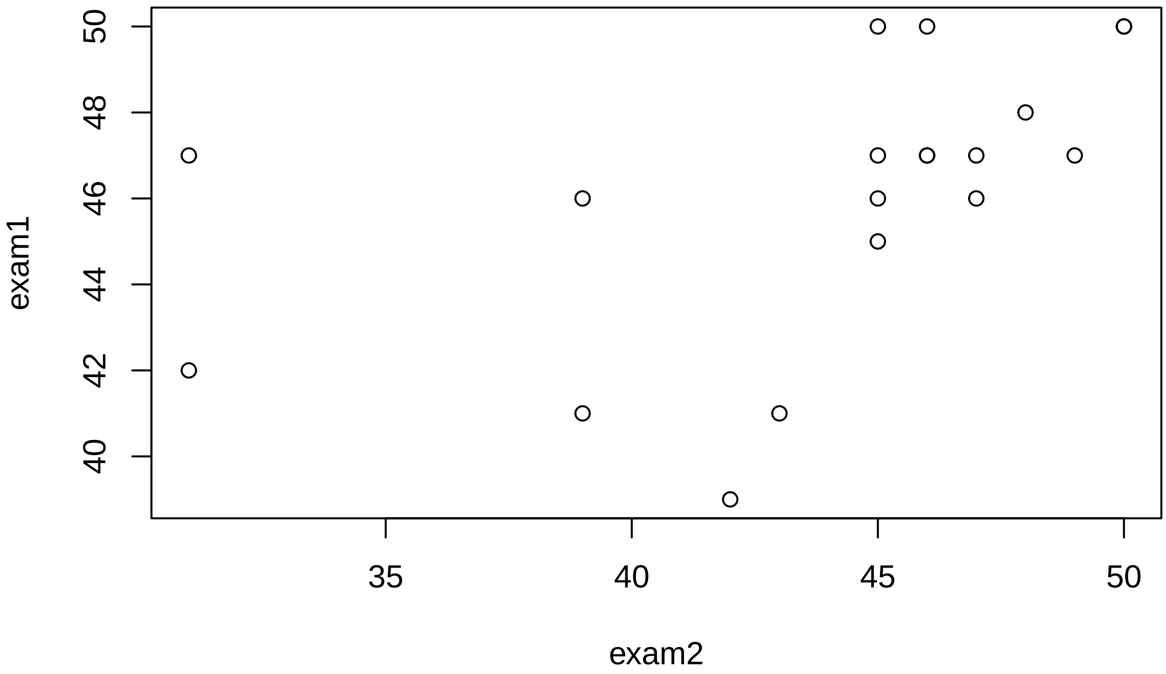
```
plot(exams$exam2 ~ exams$exam1)
```



(Note that I used the tilde ~ instead of a comma in the command above. Using a comma also works, but I like the tilde since it indicates we want to investigate the relationship between exam 1 and exam 2 scores.)

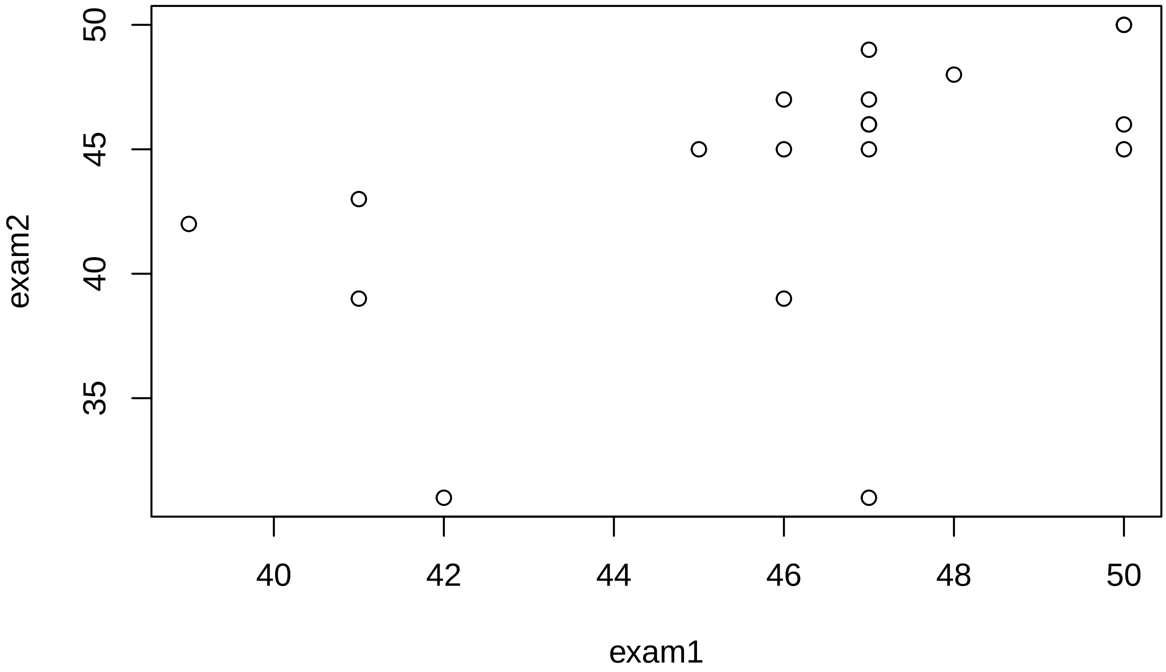
You might be tired of always typing the name of the dataframe all the time—similar to using someone's full name all the time, instead of using a nickname. There is a way to avoid this:

```
with(exams, plot(exam2,exam1))
```



You can also use the `attach` command.

```
attach(exams)
plot(exam2 ~ exam1)
```



```
#help('with')
#help('attach')
```

I do NOT like this option because it can cause vectors to be overwritten. I prefer calling the dataframe everytime I make a plot.

Question: Do you get a warning message here? If so, what do you think it means?

The warning message is: "The following objects are masked from exams (pos = 3): exam1, exam2, ID, Improve"

I think it means that it could end up overwriting the original vectors because attach ends up loading new ones.

Question: can you interchange `attach` and `with`?

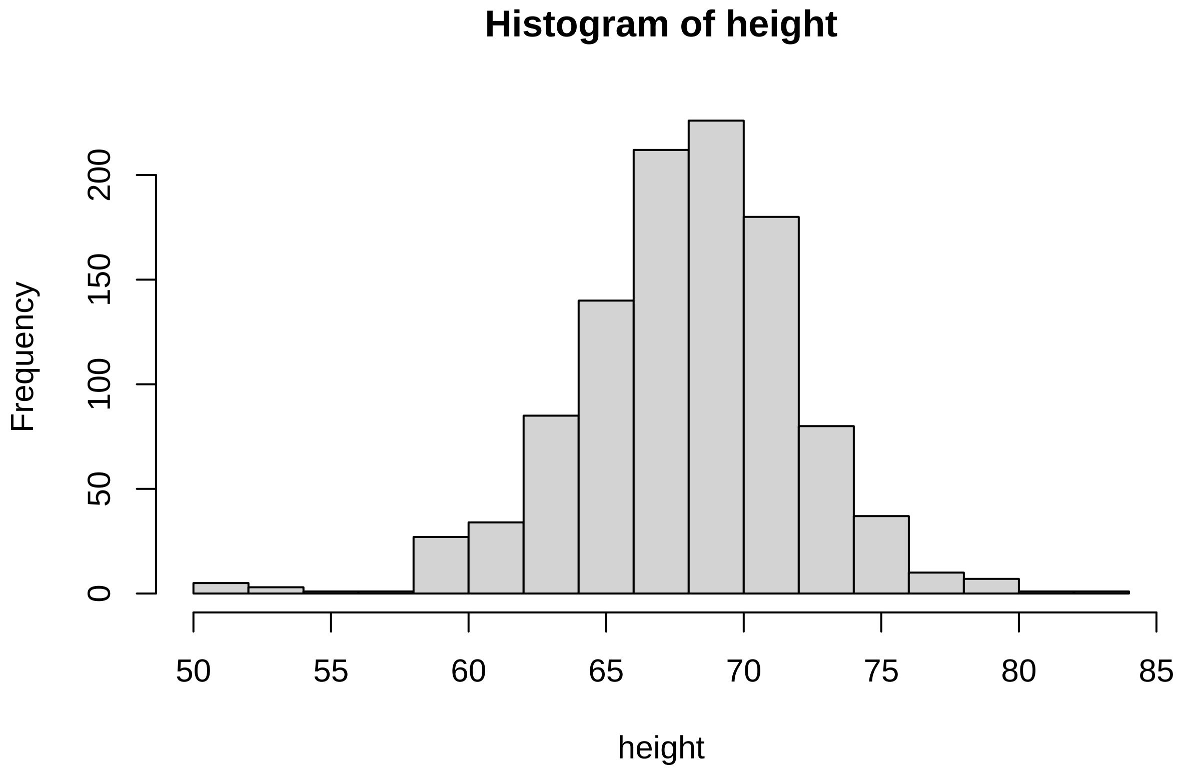
No, because the way attach and with function are different. With creates a temporary reference to the variables without masking the original variables of the data frame.

Another question: how would you describe in your own words the relationship between exam1 and exam2?

As exam1 scores increases, exam2 scores to some degree, increases as well.

We visually investigate relationships between two numeric variables by using scatterplots, like the above. When you want to see the distribution of just one variable, we usually use a histogram. Note that histograms are particularly useful when there are a lot of data. Let's use the `heights` dataframe as an example. It contains self-reported heights of 1050 individuals.

```
with(heights, hist(height))
```



How would you describe the distribution of heights?

The frequency is greatest between ~67 and ~73

If you want to compare a quantitative variable across several different groups, you can use boxplots.

If you haven't seen boxplots before, you should look it up.

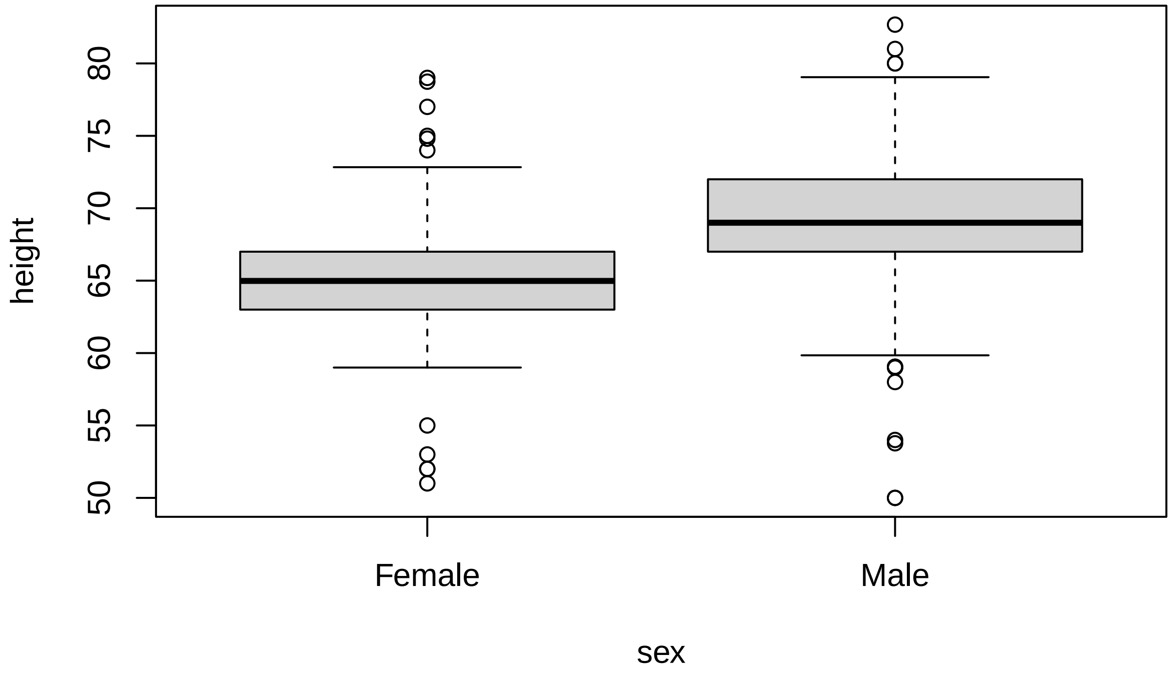
Here is an example. The `heights` dataframe also has a `sex` variable, which is of type `factor` with two levels. (Guess what they are.)

```
str(heights)
```

```
## 'data.frame':    1050 obs. of  2 variables:
##  $ sex   : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 1 1 2 ...
##  $ height: num  75 70 68 74 61 65 66 62 66 67 ...
```

We can make the boxplots:

```r
attach(heights)
boxplot(height~sex)
```



Describe what the plot above tells you about male and female heights. Be quantitative in your answer. Use the box and whiskers to help. Don't know what they represent, look it up!

1. Male height on average is taller than Female based on the IQR.
2. Males have higher maxes and mins compared to females but have more outliers below the min side.
3. Females have more outliers above max.

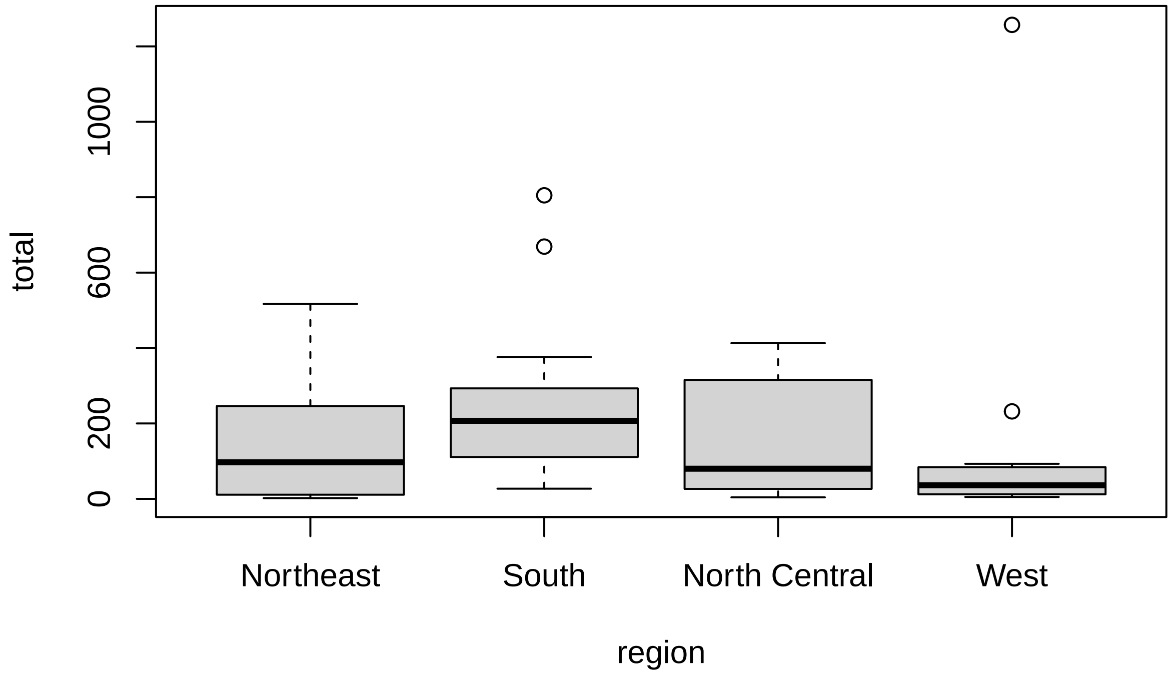
it outputs a box plot that does not represent heights based on sex, but one that does not represent anything.

Using any dataset that is available in an r package (no need to scrape data from the internet, but we will do that later!), make another boxplot using a variable that is a `factor` with at least two levels. I like sports analytics, there is a cool package called `Lahman` that has more MLB baseball data than one could ever need!

When making your boxplot, explore the help documentation and make the plot more visually appealing or clear in at least one way.

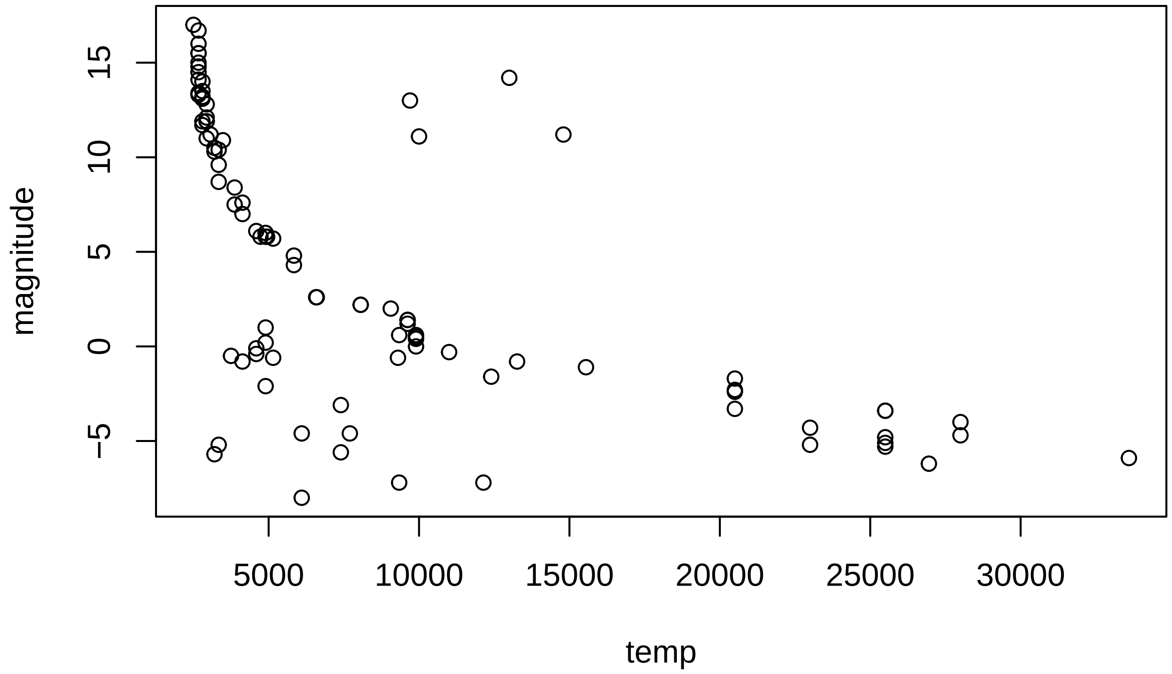
```r
# Write your boxplot code chunk here.

with(murders, boxplot(total~region))
```



Make the appropriate plot for two quantitative variables. Once again, use any readily available data you would like.

```
# Write the code for your quantitative variables plot here.
#stars
with(stars, plot(magnitude ~ temp))
```



```
#with(stars, hist(temp))
```